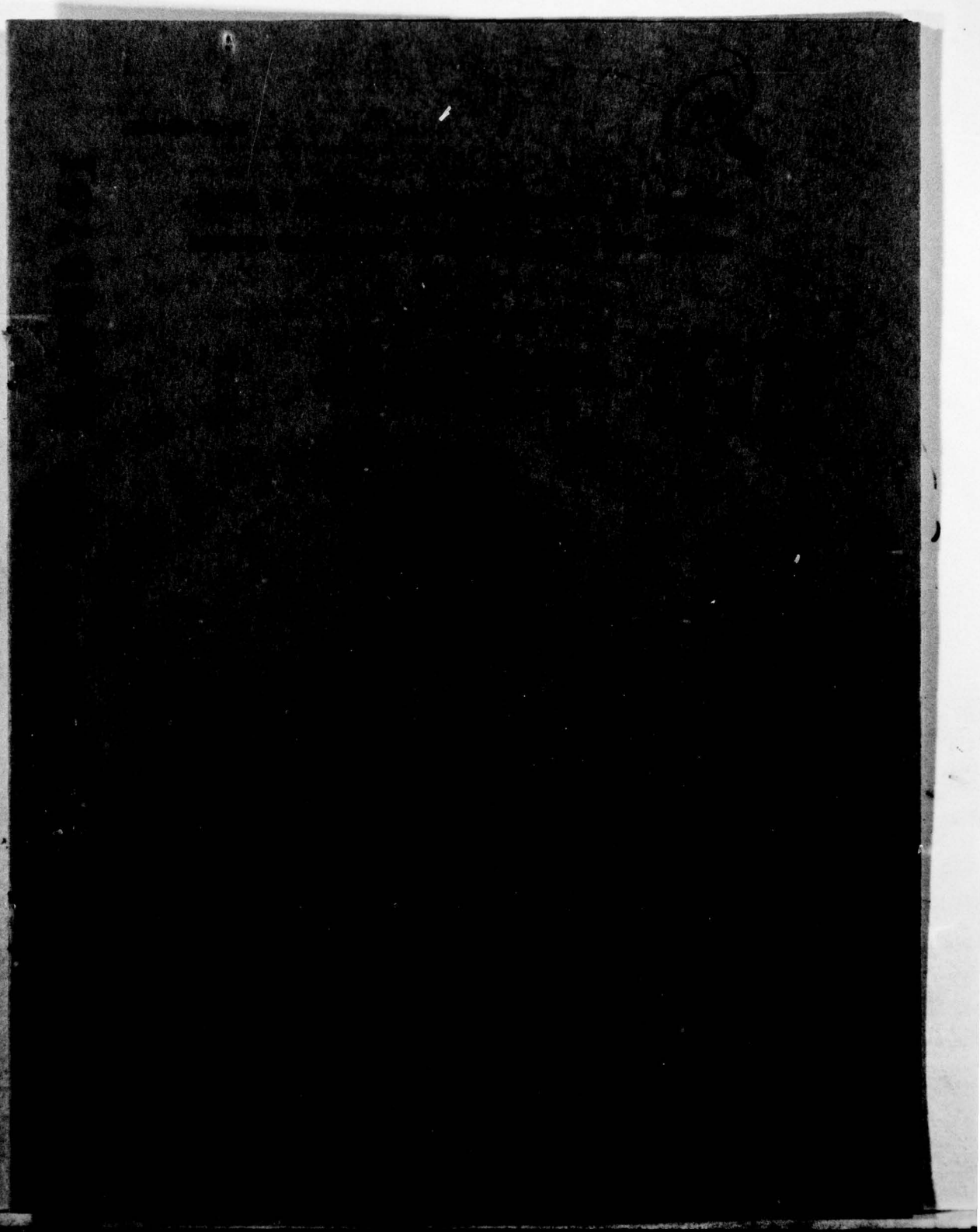
AD-A063 731 FEDERAL AVIATION ADMINISTRATION WASHINGTON D C OFFICE--ETC F/G 5/10
A METHOD TO EVALUATE PERFORMANCE RELIABILITY OF INDIVIDUAL SUBJ--ETC(U)
OCT 78 A E JENNINGS

UNCLASSIFIED FAA-AM-78-37 NL

1 OF
AD
A063731

END
DATE
FILMED
3-79
DDC



Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FAA-AM-78-37		

4. Title and Subtitle
A METHOD TO EVALUATE PERFORMANCE RELIABILITY OF INDIVIDUAL SUBJECTS IN LABORATORY RESEARCH APPLIED TO WORK SETTINGS

5. Report Date
OCTOBER 1978

6. Performing Organization Code

8. Performing Organization Report No.

7. Author(s)
ALAN E. JENNINGS

9. Performing Organization Name and Address
FAA Civil Aeromedical Institute
P.O. Box 25082
Oklahoma City, Oklahoma 73125

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

13. Type of Report and Period Covered

12. Sponsoring Agency Name and Address
Office of Aviation Medicine
Federal Aviation Administration
800 Independence Avenue, S.W.
Washington, D.C. 20591

14. Sponsoring Agency Code

15. Supplementary Notes
Work was performed under Task AM-D-78-PSY-57.

16. Abstract

This report presents a method that may be used to evaluate the reliability of performance of individual subjects, particularly in applied laboratory research. The method is based on analysis of variance of a tasks-by-subjects data matrix, with all scores standardized. If all tasks are parallel, then the average correlation among tasks is an inverse function of the within-subject variance, which may be computed for any individual subject or group of subjects. The formula for determining the relationship between within-subject variance and average correlation is developed and a method of testing the reliability of individual subjects against the general level of reliability is presented. Possible applications of the method are noted.

17. Key Words	18. Distribution Statement
Complex Performance Performance Reliability	Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	5	

A METHOD TO EVALUATE PERFORMANCE RELIABILITY OF INDIVIDUAL SUBJECTS IN LABORATORY RESEARCH APPLIED TO WORK SETTINGS

I. Introduction.

In laboratory research designed for eventual application to work settings, frequently the purpose is to be able to generalize performance of one population (say, college students or aviation cadets) on a complex laboratory task to a population that is highly selected for ability and motivation, e.g., airline pilots or air traffic controllers. When the tasks under consideration are complex, there is frequently a training phase of the study during which the subjects are familiarized with the tasks. If the aim of the research is to generalize to a population that is both highly skilled and motivated, it is often appropriate to select subjects during this training phase who can perform the test tasks at some minimum level of competence and who exhibit sufficient motivation to maintain consistently acceptable performance. This is especially important in this type of research because data collection is often very time consuming and costly, and practical considerations limit the sample size. An incompetent or unreliable subject can dramatically affect the accuracy of the results of such studies and, therefore, the appropriateness for applying research outcomes to the target population. An incompetent subject may be identified by specifying a minimum level of performance in the training phase of a study. However, especially in cases where repeated measure designs are employed with a small number of subjects, it would also be desirable to identify subjects who exhibit low reliability during training in order to eliminate such subjects from further training and testing. In such cases, grossly unreliable performance may be reasonably interpreted to indicate inadequate motivation or ability on the part of a subject. That is, a subject who attends to the task and performs adequately part of the time and at other times virtually ignores the task and performs at very poor levels will have corresponding variations in the task performance measure. Such variability of performance would not be likely (or acceptable) in the "real life" situations that are the ultimate concern of such research. If, for example, the researcher is generalizing to pilot performance, a pilot who was occasionally uninterested in the accuracy of his landing approach would be rapidly eliminated from the population of pilots, if not the population of the living. Thus, the elimination of subjects who clearly are able to perform adequately but who are unwilling or unable to maintain acceptable levels of performance may be an important factor in the generalizability of research findings.

In research designs where multiple measures of the same variable are made on the same subject (repeated measures), reliability of the measure is frequently estimated through the use of analysis of variance (1,4). The intent of such an estimate is to assess the stability of the test or to define

homogeneous subsets of test items. The present study develops a method that may be used to estimate the reliability of an individual subject's performance across successive administrations of the same task or parallel versions of the same test and identify subjects with extremely low reliabilities. Identification of such subjects is particularly useful when the sample size is small and an unreliable subject can significantly affect the validity of the research results.

II. Method.

If, in a subjects-by-measures data matrix, all within-measure variances are equal, then the average correlation (including the diagonal) (R) among the measures is equal to the sum of squares for subjects (SS_s) divided by the quantity, total sum of squares (SS_t) minus sum of squares between measures (SS_a);

$$R = SS_s/(SS_t - SS_a).$$

If within-measure variances are unequal, then R in the above expression is a function of the sum of the covariance matrix rather than the average correlation.

This average correlation among measures (R) is an estimate of reliability of the measures, if they are parallel (6, p. 61). Parallel measures are distinct measurements that measure the same thing on the same scale (6, p. 48). Therefore, the intercorrelations of parallel measures should be equal and are the upper bound on correlations with other tests (6, p. 59).

Since the purpose of this analysis is to derive an index of subject reliability rather than measure differences, all measures must be standardized within administrations. This has the effect of equalizing the within-measure variances and results in reducing the sum of squares for measures (SS_a) to zero.

Since $SS_a = 0$, $R = \frac{SS_{subj}}{SS_{total}}$. SS_{total} is equal to the sum of SS_{subj}, and the error term SS_{ws} (sum of squares within subjects). SS_{ws} is the sum of the squared deviations of test scores around the individual subject's mean test score, which is equal to the sum of squares for the subjects-by-measures interaction.

$$SS_{total} = SS_{subj} + SS_{ws} = SS_{subj} + SS_{subj \times a}$$

R, which is used as an estimate of reliability, can then be defined as an inverse function of the within-subject variance.

$$R = 1 - SS_{ws}/SS_t$$

The within-subject variance may be calculated for any subject or group of subjects and subsequently used as an index of reliability for that subject or group of subjects.

In order to test the reliability of a given subject against the overall level of reliability, the within-subject variance for a given subject (V_i) may be compared with the within-subject variance associated with scores from the remainder of the subjects (V_{-i}). Since these two variances are independent if all subjects are independent, they may be compared by use of an F ratio. A significant V_i/V_{-i} would indicate that subject i was significantly less reliable at the specific α level than the rest of the subject sample.

The calculational procedure for these tests is as follows. Assume a data matrix X_{ij} with i = 1 to N subjects and j = 1 to M measures. These measures might reasonably be repeated measures on the same task or measures from parallel forms of the same task. The scores in the data matrix would first be standardized so that all column (measure) means and variances are equal.

Let V_i equal the within-subject variance of subject i.

$$SS_{\text{within } i} = \sum_j X^2_{ij} - (\sum_j X_{ij})^2/M \quad (M = \text{number of measures})$$

$$df_{\text{within } i} = M - 1 \quad \text{so,}$$

$$V_i = SS_{\text{within } i}/df_{\text{within } i}$$

Let V_{-i} equal the within-subject variance of all subjects except i.

$$SS_{-i} = SS_{\text{within subj}} - SS_{\text{within } i}$$

$$= SS_{\text{total}} - SS_{\text{subj}} - SS_{\text{within } i}$$

$$df_{-i} = df_{\text{within subj}} - df_{\text{within } i}$$

$$= (M-1)(N-2) \qquad (N = \text{number of subjects})$$

$$V_{-i} = SS_{-i}/df_{-i}$$

Since V_i and V_{-i} are independent variances if all subjects are independent, the ratio between them is distributed as F, with (M-1) and (N-2)(M-1) degrees of freedom. A significant V_i/V_{-i} indicates that subject x is less reliable in his performance than the other subjects.

A problem in the application of this method is that it involves multiple tests, i.e., each subject is tested separately for reliability. In experimental situations where multiple comparisons are made, the Type I error rate (alpha) is much higher than the alpha level chosen for the individual tests. A straightforward solution to this problem is to use a smaller alpha value,

which takes into account the number of comparisons. A simple formula (8) for the determination of alpha resulting from multiple comparisons is: $alpha_e = 1 - (1 - alpha)^c$ where $alpha_e$ is the error rate per experiment, alpha is the error rate per comparison and c is the number of independent comparisons. Although the comparisons made in the present study are not independent, this approach will identify subjects who are extreme. A table of critical values for $alpha_e$ may be found in Jacobs (5).

In some situations, the experimenter may want to estimate the effect on R of deletion of certain subjects. This procedure is not readily amenable to significance testing but may be used to get a "feel" for the data.

R_{-i} = an estimate of the average correlation that would result if subject i were removed (assuming that for all measures, mean = 0 and s.d. = 1).

$$R_{-i} = (SS_{-i} - (\Sigma_j X_{ij})^2/MN)/(SS_{total} - (N/(N-1)\Sigma_j X_{ij}^2)$$

A comparison of R and R_{-x} ($R - R_{-x}$) may be used to provide an index of the effect on overall reliability of a given subject's scores.

III. Discussion.

The method presented here provides researchers with a tool that may be used to identify subjects whose performance on repeated measures or parallel measures is unusually inconsistent. The procedure can be used for preselection of subjects for experimental studies in human factors research in which practical considerations dictate small sample sizes.

The "prediction of predictability" is a problem that has long plagued researchers (2,3,7). Using a subject reliability index as a predictability measure is a concept that has not been applied. Of course, research utilizing this method is needed to determine its potential usefulness.

References

1. Cronbach, L. J.: Coefficient Alpha and the Internal Structure of Tests, PSYCHOMETRIKA, 16:297-334, 1951.

2. Frederikson, N., and S. D. Melville: Differential Predictability in the Use of Test Scores, EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 14:647-656, 1954.

3. Ghiselli, E. E.: The Prediction of Predictability, EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 20:3-8, 1960.

4. Hoyt, C.: Test Reliability Estimated by Analyses of Variance, PSYCHOMETRIKA, 6:153-160, 1941.

5. Jacobs, K. W.: A Table for the Determination of Experimentwise Error Rate (Alpha) From Independent Comparisons, EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 36:899-903, 1976.

6. Lord, F. M., and N. Melvin: Statistical Theories of Mental Test Scores, Reading, Massachusetts, Addison-Wesley, 1968.

7. Rock, D. A.: The Identification and Utilization of Moderator Effects in Prediction Systems, Research Bulletin 69-32, Princeton, New Jersey, Educational Testing Service, 1969.

8. Ryan, T. A.: Multiple Comparisons in Psychological Research, PSYCHOLOGICAL BULLETIN, 56:26-47, 1959.